PB84-121037

Synthesis of the Rural Model Reviews

American Meteorological Society, Boston, MA

Prepared for

Environmental Sciences Research Lab.
Research Triangle Park, NC

Nov 83

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA (Please read Instructions on the reverse before completing)		
1. REPORT NO. EPA-600/3-83-108	2.	3. RECIPIENT'S ACCESSION NO. PB84 121037
4. TITLE AND SUBTITLE SYNTHESIS OF THE RURAL MODEL REVIEWS		5. REPORT DATE November 1983 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) D. G. Fox, D. Randerson, M. E. Smith, F. D. White, and J. C. Wyngaard		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS American Meteorological Society 45 Beacon Street Boston, Massachusetts 02108		10. PROGRAM ELEMENT NO. CDTA1D/05-0279(FY-83) 11. CONTRACT/GRANT NO. 810297-01
12. SPONSORING AGENCY NAME AND ADDRESS Environmental Sciences Research Laboratory - RTP, NC Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, North Carolina 27711		13. TYPE OF REPORT AND PERIOD COVERED Final 10/82-04/83 14. SPONSORING AGENCY CODE EPA/600/09
15. SUPPLEMENTARY NOTES		

16. ABSTRACT

The Environmental Protection Agency has undertaken an independent review of eight rural diffusion models, two of which were developed by the EPA; the remaining six were submitted to the EPA for approval by outside agencies and consulting firms. In the first phase of the review process, EPA arranged with an outside contractor to calculate and tabulate a uniform set of statistics for the eight models to provide reviewers with a consistent set of measures for evaluating model performance.

Under a cooperative agreement with the EPA, the American Meteorological Society conducted the scientific review of the rural diffusion models. Seven independent reviewers evaluated each model using scientific and technical information obtained from User's Guides and the statistical performance data developed for the EPA. This report presents the results of the scientific review as summarized by the AMS Steering Committee, and contains some of the views and recommendations of the AMS Committee based on the review process and the performance evaluations.

17. KEY WORDS AND DOCUMENT ANALYSIS		
a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS (This Report) UNCLASSIFIED 20. SECURITY CLASS (This page) UNCLASSIFIED	21. NO. OF PAGES 52 22. PRICE

EPA Form 2220-1 (9-73)

PB84-121037

EPA-600/3-83-108
November 1983

SYNTHESIS OF THE RURAL MODEL REVIEWS

by

D. G. Fox, D. Randerson, M. E. Smith,
F. D. White, and J. C. Wyngaard

American Meteorological Society
45 Beacon Street
Boston, Massachusetts 02108

Cooperative Agreement No. 810297-01

Project Officer

Kenneth L. Demerjian
Meteorology and Assessment Division
Environmental Sciences Research Laboratory
Research Triangle Park, North Carolina 27711

ENVIRONMENTAL SCIENCES RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

PREFACE

From the time that the Environmental Protection Agency began using diffusion models to assess air quality issues and to set quantitative emission limits, the Agency has been criticized for officially sanctioning models that are unvalidated and that have not been impartially reviewed by a peer group of scientists. The Agency has also been criticized for failing to recognize models other than those in its own guideline documents as suitable for regulatory applications. The most comprehensive and specific recommendations with respect to these problems were voiced by the American Meteorological Society in 1981, in a document entitled "Air Quality Modeling and the Clean Air Act," and the review presented here is a direct outgrowth of this advice. The work we are reporting on is a first attempt to respond to the AMS recommendations that a comprehensive set of performance measures be developed and that both the performance measures and the details of the models be reviewed by an independent peer group.

As the title suggests, this report is primarily a summary of seven independent scientific reviews of eight rural air quality models and the evaluation of their performance. Although the report contains some of the views and recommendations of the AMS Committee, it does not express fully the broad reaction of the Committee members to the reviews or to the review process itself. The summary tends to focus on the inadequacies of the models, the overabundance of performance statistics, and the limitations of the data base. Its tone is therefore distinctly negative. We feel that a number of important lessons were learned during the conduct of this first review, and this Preface is an appropriate place to present them.

The review was undertaken in much the same way that one would solicit a formal critique of a scientific paper or proposal. Although the Committee provided ground-rules and asked certain basic questions of the individual

reviewers, there was no subsequent technical interchange between them and the Committee members. On the contrary, much as a journal editor might have done, we have taken the reviews, summarized them, and passed on their ideas in this report with as little distortion as possible. The reviewers were in no position to develop, nor were they asked to suggest, changes in the review procedure or in the performance measures. What suggestions they did make typically appeared as negative reactions to the models, the model performance, the data base and the statistics.

The Committee feels that the most important issue needing clarification is the poor correlation between observations and model predictions paired in space and time. This issue was stressed as a basic weakness by most of the reviewers, one which in their minds rendered many of the performance measures relatively meaningless. However, it should be remembered that any prediction represents an ensemble average, whereas any given observation reflects a specific realization that will almost always differ from the prediction, no matter how perfect the model and the input data. This difference represents an inherent uncertainty in air quality modeling, an uncertainty which can be quantified statistically only through some measure such as the variance. Therefore, one of the fundamental causes of the poor correlation is the inherent uncertainty in the problem. The other two basic contributors to the poor correlation - inadequacy of the data base and scientific flaws in the models - received detailed attention from the reviewers, but the inherent uncertainty of the predictions was not emphasized. We are not suggesting that the reviewers are unaware of this factor, but only that they did not emphasize it in their submissions to the Committee. In any case, the Committee feels it necessary to reiterate that on the basis of uncertainty alone one should not expect good correlation between observations and predictions paired in space and time.

Unfortunately, it is very difficult to distinguish between the inherent uncertainty and other sources of error without an excellent, comprehensive data base, and we can do little more than indicate that it must be a substantial fraction of a given prediction.

The one data base which could have made the comparisons between observations and predictions more meaningful is that recently obtained by the Electric Power Research Institute (EPRI) at the Kincaid generating station. Obtaining a complete and reliable data base was the primary objective of that project, and it is recommended that future evaluations of this type utilize this impressive set of records. The Committee did attempt to obtain the Kincaid records from EPRI, but they were not available.

The Committee does not want to leave the impression that the Kincaid study will suffice to answer all future questions regarding the performance of diffusion models in flat or gently rolling terrain. It is only one of the studies necessary to improve our knowledge of the transport and diffusion processes. Just as a single realization of field concentrations fails to establish the general uncertainty associated with a given model prediction, so also a single series of data obtained at one site is incapable of providing a general conclusion applicable to all sites.

The Committee does not believe that the set of performance measures calculated for this review exhausts the usefulness of the Clifty Creek data in assessing the modeling capabilities. The performance statistics developed by EPA and TRC closely followed the suggestions derived from the Woods Hole Workshop, but these data did not include certain details that might give considerable insight into the shortcomings of the models. For example, more detailed stratification of the comparisons by stability, wind speed, distance from the source or mixing height might reveal specific modeling

problems that could not be seen in the gross performance statistics. Similarly, a variety of graphical representations of the data might suggest modeling adjustments that would improve the quality of the predictions. The Committee feels that further investigation of this type might prove valuable, and it has agreed to explore additional analyses.

Finally, in our judgment this review of the rural models reveals a disturbing tendency on the part of some of our modelers to make their products resemble EPA Guideline models rather than try to make them fundamentally better. Despite the fact that EPA offered the model developers the opportunity to use whatever options and innovations they thought best, it seems apparent that their choices favored similarity to the Guideline models. We believe that these choices were made because the developers thought that this was the way to have their models approved for regulatory application.

The Committee strongly urges the scientific community to submit models that it considers technically better than those available today. Specific suggestions are to be found in the body of the report, and it would certainly seem that adoption of some of them would improve the soundness of the modeling structure. The Committee is not naive enough to suggest that models of better scientific calibre will necessarily result in better predictions from the regulatory standpoint, primarily because of the inherent uncertainty mentioned earlier. However, we firmly believe that better science should improve the predictions, and that it is therefore worth pursuing energetically.

Douglas G. Fox
Darryl Randerson
Maynard E. Smith
Fred D. White
John C. Wyngaard

ABSTRACT

The Environmental Protection Agency has undertaken an independent review of eight rural diffusion models, two of which were developed by the EPA; the remaining six were submitted to the EPA for approval by outside agencies and consulting firms. In the first phase of the review process, EPA arranged with an outside contractor to calculate and tabulate a uniform set of statistics for the eight models to provide reviewers with a consistent set of measues for evaluating model performance.

Under a cooperative agreement with the EPA, the American Meteorological Society conducted the scientific review of the rural diffusion models. Seven independent reviewers evaluated each model using scientific and technical information obtained from User's Guides and the statistical performance data developed for the EPA. This report presents the results of the scientific review as summarized by the AMS Steering Committee, and contains some of the views and recommendations of the AMS Committee based on the review process and the performance evaluations.

CONTENTS

Preface		iii
Abstract		vii
I.	EXECUTIVE SUMMARY	1
	A. The Models are Quite Similar	4
	B. The Models Do Not Reflect Current Scientific Capability	4
	C. The Data Base Was Inadequate	5
	D. The Models Showed Little Predictive Skill	5
	E. Conclusions of the AMS Committee	6
II.	BACKGROUND	8
III.	MODELS AND MODELING CONCEPTS	14
	A. Major Scientific Elements of the Models	15
	B. Use of a Single Modeling System	21
	C. Reason for the Unanimity About Modeling Concepts	22
	D. Possible Conclusions About the Models	22
IV.	MODEL PERFORMANCE	23
	A. The Correlation Between Observations and Predictions is Very Low	23
	B. There Is No Clear Superiority of One Model Over Another	24
V.	LIMITATIONS AND WEAKNESSES OF THE STATISTICAL STUDY	25
	A. A Single Set of Field Data Is Not Definitive	25
	B. The Statistics Were Too Elaborate and Redundant	25
	C. The Evaluation Does Not Display the Full Capabilities of the Models	26
VI.	EXTERNAL CAUSES FOR THE POOR MODEL PERFORMANCE	27
	A. Wind Direction and Wind Speed Are Uncertain	27
	B. Stability At Plume height May Be Misrepresented	28
	C. Height of the Mixed Layer Is Virtually Unknown	28

Preceding page blank

		Page
VII.	REGULATORY APPLICATIONS	29
VIII.	RECOMMENDATIONS REGARDING STATISTICAL PERFORMANCE MEASURES	30
	A. Reasons Why the Analysis Is Inadequate	30
	B. Recommendations	31
	C. Future Reviews	34
References		37
AMS Protocol		39

SYNTHESIS OF THE RURAL MODEL REVIEWS

I. EXECUTIVE SUMMARY

In a cooperative agreement with the Environmental Protection Agency, the American Meteorological Society has conducted a scientific review of a set of rural diffusion models. Two of these models were developed by EPA and the others were submitted to EPA for approval by outside agencies and consulting firms. Seven reviewers contributed to this project, and we, the AMS Committee, are most appreciative of the thorough, imaginative work they completed.

There were three phases of the review process. First, based primarily on the recommendations of the Conference held in Woods Hole, MA, (Fox 1981), EPA arranged with TRC Environmental Consultants, Inc., to calculate and tabulate a uniform set of statistics for all the models, to provide reviewers with a consistent set of measures for evaluating model performance. Second, a scientific evaluation of each model was prepared independently by each of the seven reviewers. They used both the scientific and technical information obtained from the User's Guides and the statistical data developed by TRC. Third, the AMS Steering Committee took the seven reviews and summarized them in this final report.

Many of the models submitted had a variety of options representing different assumptions and calculation procedures. Each developer was allowed to choose the options in his model that he thought best represented the problem.

The data base consisted of a two-year period of SO_2, plant emission and meteorological data developed by Environmental Research and Technology, Inc. under contract to the American Electric Power Service Corporation. The site was the Clifty Creek generating station in southern Indiana. These data, although excellent in themselves, had to be supplemented by offsite information from the National Weather Service in order to derive some of the input data required by the models. The following measurements were used:

Source

Clifty Creek Generating Station, three 208-m stacks. SO_2 emission rate calculated hourly.

Air Quality Data

Hourly average SO_2 concentrations from a six-monitor network ranging from 3 to 15 km from the source.

Meteorological Data

Wind: local 60-m tower

Temperature: 10-m level on local tower

Stability: Calculated by Turner method from Cincinnati NWS Station

Mixing Height: Developed from CRSTER preprocessor, based on Dayton radiosonde observations and Cincinnati surface data

This summary covers the scientific evaluation, but does not include the statistical performance measures. The latter will be published separately by EPA and TRC.

Before summarizing the main points of the reviewers' comments, a brief discussion of reasonable expectations with respect to model performance is in order. This is particularly appropriate in this summary, since both the statistics and the reviews provide a distinctly negative tone.

Dispersion models predict the ensemble average (i.e. the most likely) dispersion. However, dispersion measurements, repeated under the same mean conditions, differ from one realization to another because of inevitable and unresolvable differences in the details of both initial conditions and the dispersion process itself. Predicted and measured dispersion will therefore necessarily be different. The difference or scatter will never vanish, but presumably it can be decreased by better input data, better model physics and better calculation techniques. Even a perfect model with ideal input data will not agree with data in individual realizations.

The reviewers for the most part took a more fundamental position, namely that a good model should show a good

correlation with observations paired in space and time, even though this expectation would probably be somewhat unreasonable even with good input data. It is certainly unreasonable with the inadequate input data available for this study.

Therefore, the AMS Committee does not believe that the scientific community has necessarily failed to provide suitable models to represent rural dispersion, but rather that it is impossible to determine at present whether it has or not.

The main points stressed by the reviewers may be summarized as follows:

A. <u>The Models are Quite Similar</u>

Both in concept and performance the models are quite similar to each other. Furthermore, the options chosen by the developers for this evaluation tended to emphasize similarity with the approved EPA models, CRSTER and MPTER.

The statistics did reveal some differences among the models, but these variations are neither consistent nor important compared with the overall poor performance.

B. <u>The Models Do Not Reflect Current Scientific Capability</u>

The reviewers were nearly unanimous in their contention that the models do not reflect the most modern and appropriate scientific thinking. The most serious criticisms involve:

1. the Pasquill-Gifford diffusion parameters used in all but one of the models are inappropriate for many applications, especially for the tall stacks modeled in this evaluation;

2. the failure to employ recent developments in convective scaling in unstable conditions;

3. the crude treatment of plume behavior with respect to the inversion capping the mixed layer, and;

4. the speculative adjustment of the equations for source-terrain height differences.

C. The Data Base was Inadequate

Despite the fact that the quality of the data base chosen for the statistical study was excellent, not all of the relevant parameters were measured. In fact, it is virtually certain that deficiencies in the data would have made it impossible to identify even a perfect model. The main problems involved the lack of suitable information concerning the wind and turbulence at and above stack height, and the uncertainty regarding the depth of the mixed layer. All of this information was inferred from remote surface and upper air measurements using the CRSTER pre-processor system.

D. The Models Showed Litte Predictive Skill

None of the eight models showed much skill in predicting the measured SO_2 concentrations at the

same location and time. On the contrary, in the space-time pairing the predictions explained only about 10% of the variance of the observations on one-hour, three-hour and twenty-four-hour time scales.*

Five of the seven reviewers felt that this lack of fundamental correspondence between predictions and measured concentrations rendered the remainder of the statistical package almost meaningless.

E. Conclusions of the AMS Committee

1. Acceptability of the Models

Speaking for the AMS Committee, we believe that these models perform similarily, and that there is no reasonable basis for choice among them.

2. Development of Suitable Validation Data

It is apparent, as it has been in past attempts to validate models of this type, that comprehensive data are not available. EPA should devote vigorous effort to foster the development of such data, both within its own program and by the encouragement of others. It should be noted that it is unlikely that a single study, such as the

*The AMS Committee members noted, however, that the comparison between observed and predicted frequency distributions compared well in the upper percentiles, and that good space-time comparisons should not necessarily be expected.

EPRI project at Kincaid will provide sufficient information. The EPRI study should be part of a much larger effort, involving a number of other sites.

3. Modeling Innovations Should be Encouraged

Whether justified or not, there is a strong and pervasive impression among the reviewers that EPA tends to discourage models that do not bear a very close resemblance to the CRSTER - MPTER systems. Modeling innovations should be actively encouraged by all concerned.

4. Suggestions for Future Reviews

A detailed list of recommendations for future reviews appears at the end of this report, but the Committee feels that future reviews will be more meaningful only if:

a. they can be based on suitable meteorological and air quality data;

b. the statistical evaluation can be reduced to a reasonably digestible package, and;

c. the available models can be tested fully enough to determine whether available modeling options and innovations have merit.

II. BACKGROUND

The Clean Air Act passed in 1970 produced strong motivation for the application of mathematical models to air quality problems, and the use of models was specifically mandated in the 1977 amendments. As regulatory procedures developed after passage of these acts, models became a major factor in enormously expensive and important decisions. Precise limits on emissions, and even decisions about acceptability of a new source in a given area, were often based on numbers provided by an EPA-approved set of diffusion models.

Since September 1979, the American Meteorological Society (AMS) has been involved with the Enviromental Protection Agency (EPA) through a cooperative agreement, under which the AMS has provided expertise and assistance in evaluating technical aspects of air quality modeling. In a report to EPA, entitled Air Quality Modeling and the Clean Air Act (AMS, 1981) it was recommended that EPA conduct a scientific review of all models currently listed in the modeling Guideline and those being considered for regulatory applications. Further, the AMS recommended that this review should be based in part on a statistical evaluation using the performance measures developed previously in an AMS workshop (Fox, 1981).

In June 1981, following a public call for models in the March 27, 1980 issue of the Federal Register, EPA formally asked the AMS to undertake a scientific review of ten "rural" point source models which they were considering. In August 1981, the AMS agreed to this request, and the

purpose of this document is to summarize the scientific reviews which have been completed.

The ten models submitted for the review were:

MPTER/CRSTER	Environmental Protection Agency
PLUME 5	Pacific Gas & Electric
MPSDM	Environmental Research & Technology
COMPTER	Alabama Air Pollution Control Commission
SCRSTER	Southern Company Services
3141, 4141	Enviroplan
TEM-8A	Texas Air Pollution Control Board
MULTIMAX	Shell Oil

Since MPTER, CRSTER, and PLUME 5 will all produce identical results if the sources are in the same location, MPTER was run in the CRSTER mode, and the list of models was reduced to eight.

The AMS selected a small steering committee consisting of D. Fox, D. Randerson, M. Smith, F. White and J. Wyngaard to organize this effort. This group developed a "protocol" and a list of review questions, both of which are included in the Appendix of this report.

Seven reviewers who were considered expert in the "rural" modeling field were selected and these individuals, listed below, have submitted reports to the steering committee.

Mr. James F. Bowers, Jr.
H. E. Cramer Co., Inc.
Salt Lake City, Utah

Dr. Gabriel F. Csanady
Woods Hole Oceanographic Institute
Woods Hole, Massachusetts

Dr. Conrad J. Mason
Univ. of Michigan
Ann Arbor, Michigan

Dr. Robert N. Meroney
Colorado State Univ.
Ft. Collins, Colorado

Dr. Michael T. Mills
Teknekron, Inc.
Concord, Massachusetts

Dr. Allen H. Weber
Certified Consulting Meteorologist
Aiken, South Carolina

Dr. Jeffrey C. Weil
Martin Marrietta Corporation
Baltimore, Maryland

While the selection of the reviewers was being conducted, EPA contracted with TRC Environmental Consultants, Inc. to develop a data base for the study, and to compute a set of performance measure statistics comparing the modeling calculations with the field data. Insofar as possible, TRC was expected to duplicate the statistics suggested in the Woods Hole Workshop (Fox, 1981).

The TRC computations will be released in a separate document, and there is no need to restate all of the

assumptions or to summarize the statistics in this report, but the data base is of interest. The SO_2 and meteorological data measured at the Clifty Creek Power Plant for the years 1975 and 1976 were selected as the most suitable part of the American Electric Power records for evaluating rural models. The Clifty Creek Plant, operated by the Indiana-Kentucky Electric Corporation, is a coal-fired, base-load facility located along the Ohio River in southern Indiana. Three, 208-meter stacks were used throughout the study period to vent plant emissions. Terrain surrounding the plant consists of low ridges and rolling hills, but hill and ridge top elevations do not exceed stack height.

Hourly air quality data were acquired from a six-station network of continuous SO_2 monitors (Meloy) located in southern Indiana and northern Kentucky and ranging from about 3 to 15 kilometers from the Clifty Creek Plant. These data, together with onsite winds and temperatures were combined with National Weather Service surface and radiosonde measurements, obtained from stations 90 and 160 km away respectively, to form the basic data set.

During the course of the study, a question arose concerning possible bias of the SO_2 data due to CO_2 interference in the Meloy sampling system. This possibility was carefully restudied by ERT, who had installed and operated the monitoring network, and the Committee is satisfied that the bulk of the data were unbiased, and that the small portion which might have been affected shows no evidence of the slightly lower readings that might have been expected.

It was agreed among the AMS Committee, EPA and TRC that a wide range of performance statistics be calculated on one-hour, three-hour and twenty-four hour time scales. These statistics and the User's Guides prepared by the developers were sent to each of the reviewers in April, 1982.

It is important to explain why the report on the TRC computations contains an Appendix summarizing a complete rerun of the MPSDM statistics. During the course of the evaluation it was found that the original set of statistics calculated for this model misrepresented the model performance because of misinterpretations of input units and internal code procedures unique to that particular model. At the request of the AMS Committee, EPA reran the study for MPSDM. This rerun was finished after the reviewers had completed their work and their conclusions were based on the original erroneous data. The AMS Committee studied the revised statistics and decided that these new data could not have a significant effect on the conclusions presented in this synthesis. We therefore declined to ask the reviewers to make a reassessment of their positions.

The committee assigned two of its members, White and Smith, to manage the details of the review task. The full steering committee has participated in the preparation of this summary of the individual reports. In addition, each of the scientific reviewers has had the opportunity to check this summary to be certain that his views are adequately reflected. Finally, the Executive Director of the American Meteorological Society has reviewed the document.

In accordance with prior agreement between the AMS Committee and EPA, the individual reviews themselves are not appended to this summary.

In developing this summary, we found it difficult to organize the material along clear-cut lines. This was particularly true since the reviewers were under no particular constraints to distinguish among fundamental modeling concepts and the statistical performance study. However, we have attempted to segregate our discussion along these lines to make the document easier to follow.

III. MODELS AND MODELING CONCEPTS

The comments in this section were developed directly from the reviews, rather than from the ideas of the Committee members.

The AMS Committee asked the panel of reviewers to consider the models from several different standpoints. The first had to do with the physical concepts involved. The purpose of this part of the review was to highlight scientific strengths and weaknesses which might be important in applications other than those reflected in this evaluation, and to avoid being overly impressed by a model which fortuitously performed well in this particular context but might not in another. Secondly, the Committee was interested in whether the reviewers considered the models to be scientifically up to date, reflecting the best that the scientific community has to offer.

There was virtual unanimity about the scientific status of the models. The group felt that they were reviewing a single, physcially incomplete, Gaussian model with slight variations, rather than eight substantially different models. One of the reviewers noted that the entire study could have been conducted with a single "supermodel" which had the Gaussian core and a series of options or embellishments. A number of the models contained options that might have made a difference in their performance, but the statistical review tended to focus on those options which essentially duplicate the EPA CRSTER-MPTER system. The modelers themselves selected these options.

Most of the reviewers feel that these models do not represent "state of the science" in diffusion modeling, and all had serious criticisms of specific parts of the modeling systems. The reviewers believe that important changes need to be made to bring the modeling into accord with current knowledge and capabilities.

A. Major Scientific Elements of the Models

1. Gaussian Plume Model

All eight models are based on the Gaussian plume concept that has been widely used for decades. The four reviewers commenting on this point believe that this algorithm does not take advantage of recent improvements in our knowledge of the structure of the planetary boundary layer, and that it therefore does not represent the best that the scientific community has to offer. Several expressed surprise or disappointment that there was so little scientific variability among the models.

Criticism specifically centered on failure to include improvements in the modeling of convective situations. Several of the reviewers believe that significant advances have been made in our understanding of the planetary boundary layer, especially in the numerical and laboratory modeling, and in the use of convective scaling parameters (Deardorff and Willis, 1978 and 1981: Lamb, 1979; Weil and Brower, 1982). Furthermore, it was pointed out, quite correctly,

that these light wind, convective situations are often involved in the short-term maximum concentrations that are of particular interest to the regulators. The comments of our reviewers suggest that scientific support exists for the development of improved modeling of these situations as an alternative to the current CRSTER-MPTER practice.

The reviewers were also concerned about the failure of the Gaussian modeling systems to deal in a sound scientific manner with calm and near-calm conditions.* These steady-state models are not applicable to such conditions; yet they are forced to include them by highly questionable procedures in which the wind direction is assumed to be persistent and the speed is set at an arbitrary value. One reviewer pointed out that these are the very situations in which the straight-line transport from source to receptor assumed in the Gaussian models fails entirely.

2. Pasquill-Gifford Diffusion Coefficients

The time-honored Pasquill-Gifford (P-G) diffusion coefficients are certainly not favorites of this set of reviewers. They were nearly unanimous in suggesting that different coefficients be used in models, particularly if the models are to be applied to tall stacks as they were in this review.

*In this particular study, this inadequacy of the models was not important because the onsite data contained no calms.

They reiterated earlier criticisms that the Pasquill-Gifford expressions are based on small-scale, short-term, ground-level data, have not been shown to apply to elevated sources.

There was also criticism of the stratification induced by the use of any classification system of this type. The point is that the atmosphere is a continuum of motion and that arbitrary stratification can induce errors. The reviewers pointed out that there are methods of relating the diffusion directly to wind fluctuations, and that these should be used instead of a discrete system.

One of the models, MPSDM, uses a modification of the Brookhaven diffusion categories rather than those of Pasquill-Gifford, a system which should be more representative of diffusion from elevated sources, and which was developed for hourly estimates. However, this system retains the undesirable stratification of diffusion into categories, and most reviewers would prefer the direct use of turbulence measurements.

The models differ considerably from one another in the adjustment of the lateral diffusion coefficient for averaging time. Some use the P-G coefficients directly to represent hourly averages, while others use multipliers to account for the enhanced dispersion in periods greater than a

few minutes. Clearly, the reviewers were disturbed about these averaging time adjustments, and felt that there was little sound theory or firm data behind the choices.

3. Stability Classification System

The particular method used in this evaluation is in disfavor with most of the reviewers. The main complaints were that the Pasquill-Turner system is based on surface observations and fails to take account of the variation of turbulent properties with height, and that the system has an extremely strong bias toward neutral conditions, which are probably much less frequent than either stable or unstable conditions at the height of the taller stacks. It was pointed out that there are other means of deriving the stability classification, as for example, by using the ratio of the Monin-Obukhov length to the mixed layer depth, or the ratio of the friction velocity to the convective velocity scale (during unstable conditions).

4. Briggs Plume Rise

All of the models used plume rise formulas developed by Briggs. Generally, the reviewers considered this approach to be reasonable, but there was concern about proper treatment of dispersion induced by the plume motion, as well as the interaction between the plumes and the

top of the mixed layer. One reviewer would have preferred that the modeling systems include Briggs' more recent and up-to-date work (1975), for plume rise in neutral and convective conditions.

5. Mixing Heights and Plume Penetration

The penetration of buoyant plumes into elevated inversion layers is basically a question of modeling theory and concepts, but it is so intertwined with the assumptions and procedures used to estimate the mixing height itself that it is difficult to discuss the penetration problem separately.

Most of the reviewers are dissatisfied with the "all or none" concept with respect to plume penetration. The assumption is made that if the calculated effective height of the plume exceeds the base of the mixed layer, it is trapped in the upper stable layer and has no further effect on ground-level concentrations. In fact, the behavior of a plume under these circumstances depends upon the residual buoyancy it has when it reaches the elevated inversion, upon the stength of the inversion, and eventually, upon changes in the mixed layer itself.

The reviewers were very critical of the way in which the CRSTER preprocessor system derives mixing heights for the hourly input data. This system takes the twice daily mixing heights and the hourly surface data from NWS stations, and

converts the data to hourly estimates of mixing heights. This system is considered scientifically indefensible by the reviewers, and they point out that a better interpretation scheme could be based on a dynamic model using surface heat flux, measured lapse rates and temperature versus time functions.

It seems evident that the concern of the reviewers was closely related to the fact that the models often predicted no surface concentrations whatever when very high concentrations were actually observed. They assumed that at least some of these cases occurred because of poor estimates of mixing heights.

6. Terrain Corrections

The systems of correcting for terrain elevation were considered crude and unproven from any scientific standpoint. In this instance, it was not apparent that the reviewers felt that the models were lagging behind the scientific capabilities, because we are still learning how to make terrain adjustments intelligently. It was more a sense of dissatisfaction because the modelers are pretending that they know how to make such an adjustment.

7. Receptor Arrays

Only one of the reviewers went into detail on the adequacy (or lack of it) of the receptor arrays

available in the different modeling systems. He concluded that one should have at least 400 receptor points available to do a credible job of isolating the maximum effects of a source, or source configuration, and he finds only three of the models, SCSTER, Plume 5 and TEM acceptable on this basis. This subject was not discussed by the other reviewers.

It would appear that the newer data from the EPRI studies near the Kincaid Plant may shed some light on the degree of receptor detail required.

B. Use of a Single Modeling System

None of the ten models takes account of the fact that one should expect significant differences between the plume behavior from low and high sources. In their discussion, the reviewers seemed more concerned about the differences in diffusion conditions at various elevations, rather than problems related to building downwash or to other low-level aerodynamic phenomena.

To some of the reviewers this lack of distinction seemed an oversimplification. Perhaps the concern would be lessened if the management of the input data in the preprocessor system took account of the fact that diffusion conditions vary considerably with height, even within the mixed layer. Different sets of sigma curves for low and high sources would be one possible solution.

C. Reasons for the Unanimity About Modeling Concepts

At first, the similarities among the eight models struck reviewers, as well as the AMS committee members, as a bit surprising. Despite the fact that the model developers were allowed to choose the options in their programs which seemed best suited to the evaluation, their selections favored similarity with the EPA models.

D. Possible Conclusions About the Models

On the basis of these statistics and the peer reviews it would be difficult to conclude that any of the models would perform significantly better than any other.

IV. MODEL PERFORMANCE

The reviewers devised many ingenious techniques for digesting the voluminous statistics generated by TRC, each trying in his own way to decide which model most faithfully matched the field observations. It is impractical to summarize the methods they used in this document, but their conclusions are easily summarized: there is no clear superiority of any model over the others. One model may perform best in one comparison, but that same model may also be the worst in another. One can distinguish differences among the models, but these differences become meaningless when viewed in the context of the overall poor performance. Because of this performance, the committee has chosen not to reproduce any of the reviewers' comparison tables in this summary. We see little merit in reiterating the slight superiorities and inferiorities that appear in the complete set of statistics. Readers who are sufficiently interested can study the details in the EPA-TRC publication.

From a very general viewpoint, there are probably no more than two key statements to be made about the model performance as revealed in this study.

A. The Correlation Between Observations and Predictions is Very Low

The correlation coefficients linking observations and predictions paired in space and time over short time periods are near zero, with none of the models explaining more than 10% of the variance. And, while

it is true that the current regulatory practice does not focus on the space and time pairing, many of the reviewers believe that this lack of fundamental correspondence indicates either a) that the models are deficient or b) that the data and the manner in which they were used in the evaluation fail to reflect the capabilities of the models. Five of the seven reviewers feel that if the basic correlation between observations and predictions paired in space and time is poor, the other statistics are probably of little value.

B. <u>There is No Clear Superiority of One Model Over Another</u>

Although the comparative tables and graphs prepared by the reviewers were interesting and informative, they revealed no clear superiority of any one model over the others.

V. LIMITATIONS AND WEAKNESSES OF THE STATISTICAL STUDY

Although the reviewers recognized the difficulties and effort entailed in a study of this type, they did point out some important shortcomings that should be kept in mind in future undertakings.

A. A Single Set of Field Data Is Not Definitive

There was consensus that basing conclusions about the capabilities of the models on a single set of field data is not good practice. Although the reviewers were not always explicit, it seems clear that they would have been dubious even if the results had been much more encouraging because the study was restricted to a single data set.

It is also important to note that this statistical exercise was a tall-stack evaluation, and there is no way to estimate how the models might perform with low-level sources.

B. The Statistics Were Too Elaborate and Redundant

Although it was clear to the AMS Committee, it may not have been clear to the reviewers that the EPA-TRC study was an attempt to develop the complete statistical package suggested in the Woods Hole Conference (Fox, 1981). A somewhat simpler package of performance measures prepared for a larger set of model options and additional field data, preferably from other sites, would have been more valuable.

C. The Evaluation Does Not Display the Full Capabilities of the Models

As noted earlier, the statistical studies were completed using only those options in the various modeling systems that the model developers selected. Although among them these options permitted a degree of variability in the treatment of various phenomena, the reviewers would have preferred to see a wider range of options employed.

There is no way to tell whether employment of other options would have produced significant gradation among the models. Most of the reviewers would have preferred to see a wider range of options included.

VI. EXTERNAL CAUSES FOR THE POOR MODEL PERFORMANCE

The reviewers were unanimous in recognizing that serious inadequacies in the input data may account for a large part of the disappointing model performance. In fact, it is probably true that one might fail to recognize an excellent model because of these limitations.*

It should be remembered in the following discussion that some of the defects in the input data could not be remedied in any historical data set such as the Clifty Creek records. Despite the fact that the SO_2 measurements and basic meteorological data were of excellent quality, there were no onsite measurements of certain key factors such as stack height winds and turbulence and the depth of the mixed layer. These ideal data did not exist at any site and they had to be approximated in order to conduct any sort of analysis. There are criticisms, however, of the way in which some of the data were approximated.

A. Wind Direction and Wind Speed Data are Uncertain

Without doubt, one of the most serious limitations is that nothing is known directly of the winds at and above stack height. Although the data used in this analysis came from an onsite 60-meter instrument, it is virtually certain that these data often did not reflect the true state of affairs at higher

*The AMS Committee recognizes the inconsistency of the reviewers in stressing both the weaknesses of the input data and the poor performance of the models, but both points were emphasized.

elevations. This weakness alone could easily account for very large discrepancies in the apparent travel of the plume, and in the estimates of plume rise as well.

The reviwers were also critical of the way in which the CRSTER-MPTER preprocessor system deals with calm hours, adjusting such hours to previous directions and speeds. The reviewers' comment is inappropriate to this particular study since the Clifty Creek data contained no claims, but their view is included since it would be applicable to other studies.

B. Stability at Plume Height May Be Misrepresented

The estimates of stability in this study were developed from the Pasquill-Turner system, itself no more than a rough approximation from standard surface data. These estimates are not considered representative of the stability at plume elevation.

C. Height of the Mixed Layer is Virtually Unknown

Within the framework of the existing upper air meteorological network, there is little that EPA-TRC could have done to obtain better measurements of the mixing height. They are simply constrained to use the remote radiosonde data as it is available.

However, the reviewers are discontented with the system used in the data preprocessor system to translate these radiosonde observations into hourly estimates of mixing height.

VII. REGULATORY APPLICATIONS

Having read the foregoing sections, it will come as no surprise that the reviewers are unhappy about the application of these models to regulatory problems. However, since the intent of the AMS-sponsored review was to study the scientific rather than the regulatory aspects of the rural models, these comments are not included.

VIII. RECOMMENDATIONS REGARDING STATISTICAL PERFORMANCE MEASURES

The comments in this section are in part derived from the reviews, but they essentially represent the viewpoint of the AMS Committee.

The statistical analysis conducted by EPA through TRC for the purpose of supporting the rural model review process needs considerable streamlining. The analysis provides statistics that are redundant, and the volume of information is so great that individuals are unable to absorb it. There is a real danger, furthermore, in assuming that because of the sheer volume of data the analysis is sufficient. This is not necessarily correct.

A. Reasons Why the Analysis is Inadequate

1. Only One Data Set Was Used

A single set of data was examined, a limitation raised by all of the reviewers, as well as by EPA and the AMS Steering Committee.

2. Time Versus Space

The models have had a test of their ability to simulate a time series of data, two years of hourly observations. However, the spatial distribution has been restricted to only 6 data points.

3. Data Set Inadequate

The data set was inadequate to distinguish among the performances of the various models.

B. Recommendations

We have some specific recommendations for future analyses.

1. Retain the Woods Hole Philosophy

The philosphy of the Woods Hole Workshop (Fox, 1981) was rather simple. Models should be evaluated qualitatively on their scientific merit, and quantitatively by comparsion against measured data. The quantitative comparison should be based on statistical evaluation of differences and correlation between observations and model predictions. However, formal statistical hypothesis testing is pointless because it would simply establish what we already know; that the models are incapable of duplicating individual observations. Instead, confidence intervals about the estimated statistics provide useful information. Any statistical evaluation depends critically on the quality and adequacy of the data base used. Since we have very little experience with quantitative performance, this type of evaluation should be considered experimental until sufficient information can be accumulated to improve our knowledge.

2. A Parsimonious List of Measures is Needed

Based largely on the reviewers' comments, the following list of statistical quantities is ranked according to usefulness.

a. Correlation Coefficients, Paired in Space and Time

A very valuable product is a straight correlation between observations and predictions, paired in both space and time. While difference lends itself to better-developed statistical theory, the results clearly indicate that such sophistication is unwarranted in this study. When the overall correlation, r^2 (since this measures the variance explained by the model) is below 10%, there is no need for sophisticated testing. A summary of correlation coefficients such as those in Table B of the TRC report would be useful. A nonparametric correlation coefficient (Spearman's ρ, Kendall's τ) would be useful also.

b. Bias and Variability of Difference

The measures of bias and variability are useful. The bias and the variance with t and χ^2 statistics constructed for 90-95% confidence would be sufficient.

c. Differences in Frequency Distributions

The frequency distributions are considered important in view of the current regulatory applications of the models, but the reviewers did not find them especially useful.

These three statistical comparisons should be done for two data sets: (1) the observations and predictions paired in time and space; and (2) the maximum N observations and the maximum N predictions unpaired in time and space.

This set of comparisons probably would not be sufficient to test urban and complex terrain models where spatial variations may be very significant. For these cases one could add one of the pattern measures suggested by the AMS Workshop. Also, there is considerable effort among the technical community to develop appropriate performance measures (Londergan, 1980; Buckner, 1981; Moore, et al., 1982).

3. Utilize Multiple Data Bases

The reviewers were concerned about the fact that the single data base used for this model evaluation effort was inadequate. While several partial data bases, rather than just one, could have been used, it is recognized that they too would suffer from the same inadequacies. A much more rigorous data base is being developed under the EPRI Plume Model Validation project (Hilst, 1978; Bowne et al., 1981). The AMS Committee was aware of the development of this data base and attempted to obtain it

for the evaluation.

4. Quality Control of the Computation

Inadequate quality control of the computations is as dangerous as dealing with the wrong statistics. However, quality control with so vast a set of historical data is exceedingly difficult. In future studies, it would be wise to run a limited set of computations, say a five-day series, and to review these fully with the model developers. This step would reduce the chance of significant errors or unrepresentative computations, and it would therefore increase the credibility of the results.

C. Future Reviews

In conducting any review, one often feels that the end result might have been better had the review been conducted in a different manner. This is certainly true in the case of the present evaluations of the "rural models". There has been much wasted effort, lack of clear guidance to the reviewers, poor timing in delivering the information and data to the reviewers and too much emphasis given to the preparation of the voluminous performance measure statistics.

To assist any individual or group that might be faced with the task of conducting a peer review of any additional category of models (such as urban, complex terrain, etc), we would suggest the following steps:

1. Arrange for the preparation of adequate data bases; try to have at least two sets of data so that an independent check on the statistics is possible.

2. Agree on the performance measure statistics that are really helpful in evaluating the usefulness of the model.

3. Permit the developers to update their models at as late a date as is consistent with orderly procedure. Most models undergo more or less continuous alterations.

4. Allow adequate time for the model developers to examine and evaluate brief test runs, and if necessary rerun these preliminary tests to assure that the developers agree upon the procedures to be followed.

5. Run the complete performance measure statistics.

6. Select the reviewers and send them the descriptions of the models and the performance statistics at the same time. Give them adequate time to conduct their review. Five reviewers may be sufficient for each category in the future.

7. Synthesize the reviews and make sure that the synthesized report is complete.

8. Let the reviewers and the developers comment on the synthesized report.

9. Modify the report as necessary and submit it to EPA.

The major cost in the preparation of the present review on "rural" models has been the preparation of the performance measure statistics. Most of the reviewers commented that the sheer magnitude of the performance data presented was overwhelming, and not worth the effort. We suggest that if we had it to do over again, only the performance measure statistics indicated previously should be calculated. This reduced effort should provide the necessary statistics for the reviewers to do their job.

REFERENCES

American Meteorological Society: Air Quality Modeling and the Clean Air Act: Recommendations to EPA on Dispersion Modeling for Regulatory Applications, AMS, Boston, MA, (1981).

Bowne N. E., et al.: Interim Plume Model Validation Results. EPRI EA-1788-SY. Electric Power Research Institute, Palo Alto, CA (1981).

Briggs G. A.: Some Recent Analyses of Plume Rise Observations, pp. 1029-1032 in Proceedings of the Second International Clean Air Congress. Edited by H. M. Englund and W. T. Berry, Academic Press, NY (1971).

Briggs G.A.: Plume Rise Predictions, pp. 59-111 in Lectures on Air Pollution and Environmental Impact Analyses, AMS, Boston, MA (1975).

Buckner, M. R.: Proceedings of the First SRL Model Validation Workshop. E. I. DuPont de Nemours & Co., Savannah River Laboratory, Aiken, S.C., (1981).

Fox D. G.: Judging Air Quality Model Performance. Bull. Am. Meteorol. Soc., 62:599-609, (1981).

Hilst G. R.: Plume Model Validation. EPRI EA-917-SY. Electric Power Research Institute, Palo Alto, CA (1978).

Lamb R. G.: The Effects of Release Height on Material Dispersion in the Convective Planetary Boundary Layer, pp. 27-33 in Proceedings Fourth Symposium on Turbulence, Diffusion and Air Pollution, Reno, NV, AMS, Boston, MA, (1979).

Londergan, R. J.: Validation of Plume Models, Statistical Methods and Criteria. EPRI EA-1673-SY. Electric Power Research Institute, Palo Alto, CA., (1980).

Mills M. T.: Improvements to Single Source Models, Vol. 3: Further Analysis of Modeling Results. EPA-450/3-77-0030, (1977).

Moore, G. E., Stoeckenus, T. E., Steward, D. A.: A Survey of Statistical Measures of Model Performance and Accuracy for Several Air Quality Models (Pub. No. 82180). Final Report to EPA, on contract 68-01-5845, Systems Applic., Inc., San Rafael, CA, 121 p., (1982).

Weil J. C. and Brower R. B.: The Maryland PPSP Dispersion Model for Tall Stacks. Prepared by Environmental Center, Martin Marietta Corporation, for Maryland Department of Natural Resources, Ref. No. PPSP-MP-36, (1982).

Willis G. E. and Deardorff J. W.: A Laboratory Study of Dispersion from an Elevated Source Within a Modeled Convective Planetary Boundary Layer, Atmos. Environ. 12:1305-1312, (1978).

APPENDIX

January 21, 1982

AMS PROTOCOL
for
Conducting Scientific Review
of Air Quality Models
for EPA

1. At least five to seven reviewers should be utilized for each category of models. The same reviewers should comment on all models within a given category of models (i.e., rural models, urban models, etc.). This requirement will ensure uniformity of review and allow the best comparative evaluations.

2. To understand the context of model applications, reviewers should be knowledgeable of EPA regulatory activities, recognize the issues involved and the controversey associated with models, but be scientifically capable of reviewing models.

3. Reviewers should not be considered to be reasonably objective and unbiased in their role as reviewers. It is anticipated that qualified reviewers may have various types of potential biases because of their activities within the field. Consequently, reviewers need to be chosen from different interest groups and in sufficient numbers to provide a means of balancing viewpoints. For this reason, consideration should be given to using as many as seven independent reviewers in this process.

4. Reviewers should not be associated with the development of models under review or be selected from organizations that have submitted models being evaluated. Likewise, AMS members of the Steering Committee should not participate in the review process if they have been involved in the preparation of the specific models being evaluated.

5. The AMS members of the Steering Committee will select the reviewers.

6. The review of models should be based in part on a statistical evaluation of model performance measures as recommended by AMS, using a suitable data base. In addition, case study data will be available.

7. One or two AMS member(s) of the Steering Committee will be selected to manage the review for each category of models. The manager's responsibilities will be to: contact the selected reviewers, make satisfactory financial arrangements, interface with the reviewers and model developers, and prepare a draft summary report for review by the Steering Committee.

8. The AMS members of the Steering Committee will review and approve the summary report prepared by the manager(s) before it is submitted to EPA. AMS members will have access to the individual scientific reviews during this process.

9. The names of the individual reviewers will be made public upon completion of the summary report. However, individual reviewer comments are intended for the confidential use of the AMS members of the Steering Committee. It is realized that this cannot be guaranteed. The summary of the peer review, which will be provided to EPA, will be "open" information.

Questions to be Addressed by

Reviewers of Air Quality Models

Questions on Individual Models:

1. For the applications intended for this category of models, does the model address all the source/receptor relationships that are germane?

2. To what degree are the underlying assumptions valid for a typical application?

3. Are the assumptions correctly formulated in the model?

4. Does the model use techniques that are currently state-of-the-art?

5. Are there technically better or more theoretically sound techniques?

6. Does the model make the best use of typically available data bases?

7. Are there obvious technical improvements required in the model?

8. Is the usefulness of the model consistent with the resources required to operate it?

9. What are the inherent attributes and limitations of the model?

10. Is the statistical performance of the mdoel in terms of bias, noise, variability and correlation generally acceptable or within the state-of-the-art?

11. For typical uses, can an objective statement be made about uncertainty associated with the model estimates?

12. What are the attributes and limitations of the model's performance?

13. Are there specific aspects on the application of these models in which they may produce misleading results, i.e., some models may predict fairly well at close distances but become unreliable at longer distances?

Questions of All Models within a Category (Based on Theoretical and Performance Characteristics):

1. What are the general attributes and limitations of models in this category?

2. How do models within this category compare to one another?

3. Is a specific model or models clearly superior to the other models?

4. Can these models be ranked individually, or in the groups? If so, how should they be ranked?